6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# A COMPARISON OF TECHNIQUES FOR TONE COMPENSATION IN PAYPHONE-BASED SPEECH RECOGNITION

Ben Milner and Mark Farrell,
BT Labs, Martlesham Heath, Suffolk, UK.
{ben.milner, mark.farrell}@bt.com

10.21437/Eurospeech.1999-629

## ABSTRACT

This paper compares two noise compensation techniques for increasing recognition performance on tone corrupted speech from payphones. An analysis of payphone speech is made which identifies two harmful processes - signalling tones and increased background noise. The techniques of spectral subtraction and parallel model combination are compared on this task and are shown to give some improvement. Further robustness is demonstrated using RASTA-CTM speech features.

## 1 INTRODUCTION

In several countries, including the UK, signalling tones are emitted by payphones to indicate that the call is being made from a payphone. The tones are clearly audible over the speech signal and cause a noticeable distortion. It was observed in an automated operator trial that calls from payphones were often incorrectly recognised because of this tone corruption.

Many different techniques exist for the compensation of tone-like distortion [1]. This work makes a comparative study of several of these. Both inherent robustness, through the use of improved speech parameterisations, and explicit tone compensation are tested.

## 2 ANALYSIS OF PROBLEM

Speech from a payphone has two noise components in addition to those normally associated with speech originating from an office or home phone. The signalling tones are one source of extra noise and the other is the increased environmental background noise. This occurs in many different forms such as road noise, railway station noise, human chatter, etc.

### 2.1 Payphone Signalling Tones

Historically the signalling tones were emitted by payphones to indicate to operators that the caller was using a payphone. Some standards were laid down for the frequency and duration of the tones which have been reasonably well adhered to. The most common set of signalling tones are those used in BT payphones. Figure 1 shows the spectrogram of a speech signal contaminated by payphone tones. The first high-low tone pair occurs in silence and the second in the middle of the speech signal.

A typical tone sequence comprises a 200ms high tone, 200ms of silence, a 200ms low tone and 2 seconds of silence before the cycle starts again. The frequency of the two tones is also reasonably stationary across different payphones with the high tone around 1210Hz and the low tone around 840Hz.

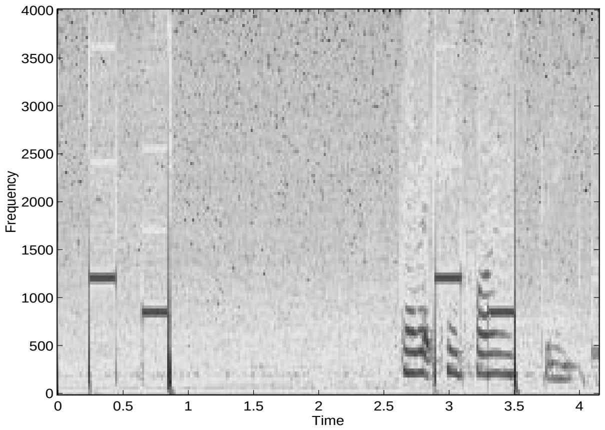

Figure 1: Spectrogram of tone corrupted speech.

This means that approximately 15% of the speech signal is corrupted by tones. As a preliminary test to see the effect of the tones on recognition performance an unconstrained monophone test was carried out. The performance was determined on approximately 6000 phonetically rich sentences from the Subscriber database [2]. Payphone tones were then added to the speech in accordance with the standards mentioned above and the monophone performance re-tested.

| | %Hits | %Insertions |
|---|---|---|
| Baseline | 55.0 | 8.9 |
| Tone corrupted | 48.1 | 16.5 |

Table 1: Unconstrained monophone accuracy.

Analysing the results shows that the addition of the tones effects recognition performance in two ways. First, tones which occur within speech are likely to cause substitution errors. This is confirmed by observing a hit rate reduction of about 15% which corresponds to the percentage of the signal corrupted by tones. Secondly, when a tone occurs in a silence period it tends to match better to a vocabulary model than the noise model. This is evident in table 1 by the doubling of insertion errors for the tone corrupted speech. This simple test shows that the payphone tones cause a significant reduction in recognition performance.

Figure 1 also shows that large impulses are produced by the tone generator when the signalling tone starts and finishes. The amplitude of this spike is often four times that of the tone amplitude and can have a duration up to 11ms. This severely whitens the frequency spectrum in this region and further distorts the signal.

## 2.2 Environmental Noise

The deployment of payphones tends to be in public places such as railways stations, pubs, hotels and on the side of the road in telephone boxes. As a result, calls made from payphones usually have much higher levels of background noise than is encountered during calls made from offices or houses.

Some preliminary experiments have shown that the conventional line-noise model was inadequate for modelling this louder background environment. To solve this a new noise model, *paynoise*, was trained from non-speech segments of payphone data. This also includes low level echo which was present.

# 3 SPECTRAL SUBTRACTION

Spectral subtraction [3] operates in a continuous manner with a noise average being updated during speech inactive periods. It is assumed that the noise is reasonably stationary. However tone distortion is very non-stationary and not well suited to conventional spectral subtraction.

Because tone corruption only occurs in short isolated bursts, a tone detector can be employed which only initiates spectral subtraction when a tone is detected. This has the advantage that spectral subtraction is not operating continuously and leaves the remaining 85% of the signal free from the processing distortions which spectral subtraction imparts on the speech.

## 3.1 Detection of Tones

The signalling tones produced by payphones are reasonably well defined. Within the duration of a tone the frequency remains stable and lasts for about 200ms. Analysis of many different payphone recordings has shown that there are five predominantly occurring tones. These are categorised in table 2.

The tone pair (*low1* and *high1*) is most commonly occurring although some payphones signal with a tone triple which uses a different set of frequencies (*low2*, *middle* and *high2*).

The detection of tones in the speech signal is a similar problem to frequency shift keying (FSK) demodulation. In particular because the phase of the tones is unknown and with five possible tone frequencies a suitable tone detector can be based on a noncoherent M-ary FSK demodulator [4].

| Tone label | Tone frequency | Occurrence |
|---|---|---|
| *Low1* | *840Hz* | Tone pair – e.g. BT payphone |
| *High1* | *1210Hz* | |
| *Low2* | *970Hz* | Tone triple – mainly privately operated. |
| *Middle* | *1230Hz* | |
| *High2* | *1530Hz* | |

Table 2: Summary of tone frequencies.

Figure 2 illustrates the system, showing a bank of five matched filters which correspond to the five tones frequencies. The envelope of the matched filter output is compared to a threshold which indicates whether a tone is present.

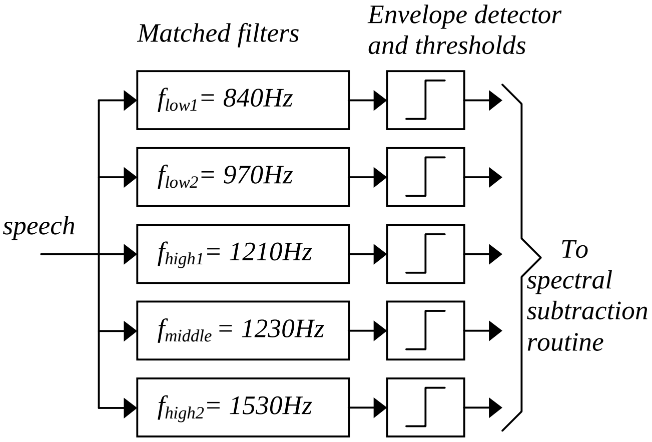

Figure 2: Bank of matched filters for tone detection.

The output of the bank of matched filters, which provide information regarding the location and frequency of tones, is passed to the spectral subtraction routine.

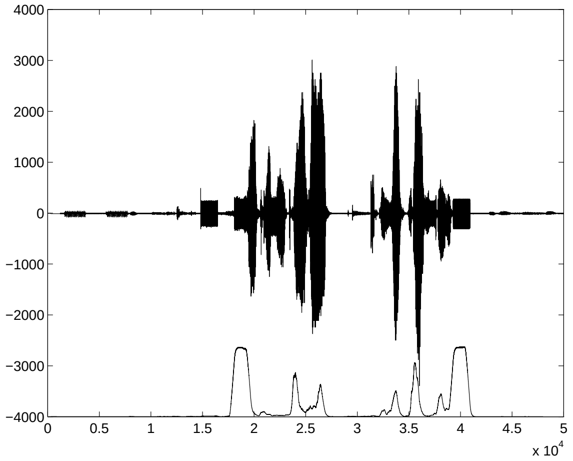

Figure 3: Speech signal and matched filter envelope.

An example of the *low1* matched filter envelope is shown in figure 3 together with the time-domain signal. The envelope clearly identifies the two instances of the low tone which occur first in speech and then in silence.

The envelope is used to identify tones using both their amplitude and duration to minimises spurious detection.

## 3.2 Tone Subtraction

To minimise processing distortions, spectral subtraction is performed in the linear mel-filterbank domain. This reduces the likelihood of a frequency bin becoming negative after subtraction and requiring post processing. Subtraction is implemented using the general form

$$\hat{X}(f) = Y(f) - N_{tone}(f) \qquad (1)$$

Where $\hat{X}(f)$ is the restored filterbank channel amplitude, $Y(f)$ is the tone corrupted channel and $N_{tone}(f)$ is the mel-filterbank representation of one of the five tones listed in table 2.

Upon implementation of the technique it was discovered that significant signal distortion was occurring in areas of low speech power. This was attributed to the fact that the effective signal to noise ratio in this region was extremely low, making signal recovery impractical.

## 3.3 Tone Interpolation

The signalling tones have a very narrow bandwidth and as such affect only two of the nineteen channels of the mel-filterbank. Because of the overlapping nature of the filterbank channels and the inherent correlation within the spectrum of a speech signal a reasonable estimate of the corrupted filterbank channels can be made from adjacent channels.

Many different interpolation strategies exist for estimating the two missing filterbank channel amplitudes. In this application a relatively simple polynomial interpolation has been used, although better performance may be obtained using some of the more sophisticated methods described in [5].

# 4 TONE MODELLING

An alternative method to filtering the payphone tones is to allow for the simultaneous modelling of tones and speech. A technique well suited to this is parallel model combination (PMC) [6].

Typically in PMC an estimate of the noise statistics is made during speech inactive periods. The vocabulary models are then inverse transformed back to a linear domain where the noise statistics can be added to the speech models. The noisy speech models are then transformed back to the original feature space.

The feature extraction transform used in this work (RASTA-CTM see section 5) is not invertable. Therefore to build the set of parallel models three sets of speech models need to be trained:

1. A set trained on undistorted speech
2. A set contaminated by *low1* tone
3. A set contaminated by *high1* tone

The models can then be combined in a manner which allows progression through the vocabulary model as well as modelling the tones. Figure 4 illustrates this showing the combination of the three sets of three state triphone models.

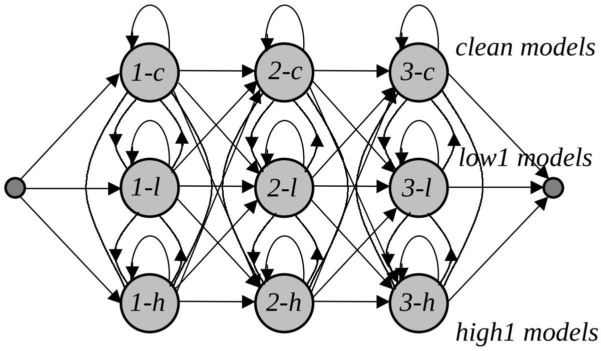

Figure 4: Parallel triphone model.

# 5 ROBUST PARAMETERISATIONS

An improved framework over differential parameters for encoding the temporal variations of speech is the RASTA-CTM cepstral-time feature matrix [8]. RASTA-CTM features are generated from a stream of RASTA filtered MFCC vectors. A block of seven RASTA-MFCC vectors, plus the log energy of the frame, are stacked together and a discrete cosine transform (DCT) computed across the stack. The resulting matrix is then truncated with the first four columns used as the speech feature. This process is illustrated in figure 5.

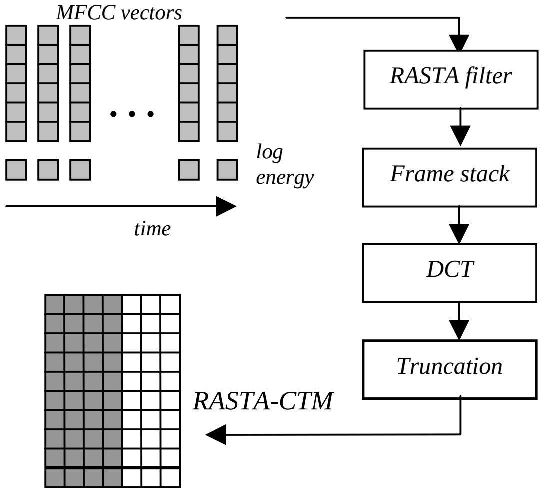

Figure 5: Generation of RASTA-CTM feature matrix.

The MFCC vectors are produced from 32ms frames of speech at 16ms intervals. The speech is passed through a 19 channel mel-filterbank with MFCCs 0 to 8 retained after the DCT. The RASTA-CTM frame rate is 16ms.

## 6 EXPERIMENTAL RESULTS

The performance of the tone compensation methods have been tested on a database collected for operator call steering applications. All calls were made from payphones and were tone contaminated.

The speech is modelled using a set of 8500 state clustered triphone models. Each triphone model comprises 3 emitting states with 12 modes per state and a diagonal covariance matrix. Noise modelling was performed using the paynoise model described in section 2.2.

Table 3 shows the baseline performance of the recognition system with the speech parameterised using both conventional MFCCs plus deltas and the RASTA-CTM features. As a comparison, recognition performance is also shown using call steering data collected from office and home based phones.

| Feature | Test Set | Word Accuracy |
|---|---|---|
| *MFCC* | *Payphone* | *35.1%* |
| *MFCC* | *Office/home* | *41.5%* |
| *RASTA-CTM* | *Payphone* | *56.6%* |
| *RASTA-CTM* | *Office/home* | *57.2%* |

Table 3: Baseline performance on payphone data.

The results show clearly that distortions made by the payphones significantly reduce recognition performance in comparison to the quieter office/home-based calls.

RASTA-CTM features improve performance considerably by incorporating inherent robustness into the recogniser.

Table 4 compares the performance of the tone compensation methods using the RASTA-CTM feature.

| | Word Accuracy (%) | Correct (%) | Insertions (%) |
|---|---|---|---|
| *Baseline* | *56.6%* | *83.3%* | *26.7%* |
| *Interpolation* | *68.1%* | *84.8%* | *16.7%* |
| *PMC* | *47.3%* | *48.4%* | *1.1%* |
| *Low/High* | *71.1%* | *81.5%* | *10.4%* |

Table 4: Tone compensation performance.

Spectral interpolation gives a 12% increase in recognition accuracy. Much of this is achieved by reducing the insertions produced by the tones.

PMC doesn't work at all on this test, although it does reduce the number of insertions to almost zero. The reduction in performance was rather surprising as tests using PMC on the subscriber database with tones added artificially did give some improvement. The final result uses just the PMC noise model (*line-noise, low1* and *high1* combined) and gives a good improvement in accuracy. This again is mainly due to a reduction of insertion errors.

Examining the recogniser output shows that many of the remaining errors were caused by impulsive noise at the word boundary. Although they are very short duration they distort a large portion of the frequency spectrum which is smeared across several frames.

## 7 CONCLUSIONS

Two harmful effects have been identified in speech collected from payphones. Combining an improved speech parameterisation with explicit tone compensation has been shown to give a significant increase in recognition performance.

A baseline MFCC system attained 35.1% word accuracy. Incorporating inherent robustness into the recogniser using the RASTA-CTM feature improved performance to 56.6%. However applying explicit compensation for the payphone tones increased performance still further to 71.1%.

Many of the errors remaining are caused by the impulsive noise. Including compensation for this is expected to give greater robustness.

## 8 ACKNOWLEDGEMENTS

The authors gratefully acknowledge the support of Frank Scahill and Peter Durston in this work.

## 9 REFERENCES

1. B.P. Milner and S.V. Vaseghi, “Noise compensation methods for hidden Markov model speech recognition in adverse environments”, IEEE Trans. Speech and Audio Processing, Jan. 1997.
2. A.D. Simons and J. Edwards, "The subscriber database", Proc. IOA, 1992.
3. S.F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. ASSP, vol. 27, no. 2, pp. 113-120, 1979.
4. S. Haykin, Digital Communications, John-Wiley, 1988.
5. S.V. Vaseghi, Advanced signal processing and digital noise reduction, John Wiley 1996.
6. M.J.F. Gales and S.J. Young, “HMM recognition in noise using parallel model combination”, Proc. Eurospeech, pp. 837-840, 1993.
7. H. Bourlard and S. Dupont, “A new ASR approach based on independent and recombination of partial frequency bands”, Proc. ICSLP, pp. 426-429, 1996.
8. B.P. Milner, “Robust speech parameterisations for telephony applications”, Workshop on Speech Recognition in Adverse Environments, Tampere 1999.